Group Assignment: Prediction of Churn of Telecommunications Customers

As you have learnt by now, predictive modeling – the process of "scoring" and targeting customers for a marketing campaign – is a significant database marketing tool and an important component of a firm's customer relationship management (CRM) effort. The promise of predictive modeling is the ability to predict what actions customers will take, thereby allowing firms to target their marketing efforts more effectively. One area of particular importance is customer "churn," in this case, customer voluntary churn, when current customers decide to take their business elsewhere or voluntarily terminate their service. Annual churn rates have been reported to be in the 20% - 40% range for telecommunication and other technology industries. This puts a premium on developing models that accurately predict which customers are most likely to churn, so proactive steps (e.g. appropriate communication and treatment programs) can be taken to prevent customers from churning. The purpose of this assignment is to figure out which method(s) works best for predicting churn, thereby enhancing our overall understanding of predictive modeling.

The Teradata Center for Customer Relationship Management at Duke University (the Center) has shared a dataset regarding the churn of telecommunications customers. The data consist of calibration and validation samples of customers from a major wireless telecommunications company. The calibration sample includes observed churn and a set of potential predictor variables. The two validation samples include the same predictor variables, but no churn variable. Your group, which has been hired as a consultant to the telecommunications company, are required to submit the predictions of likelihood to churn.

The Wireless Industry*:* Over the years, the wireless sector has been one of the fastest-growing businesses in the economy. With a unique value proposition – freedom and connectivity – the number of subscribers doubled every two years during the 90's. Wireless stocks grew as fast as those of many dot-coms, start-ups emerged everywhere, and IPO's raised record amounts of money.

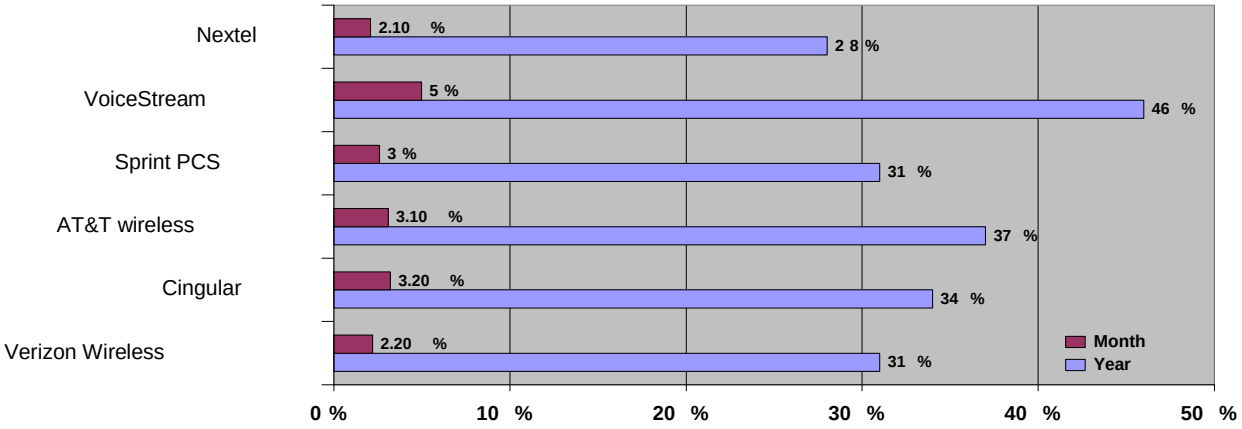
These events shaped the new telecommunications landscape as we know it today.

Industry Turmoil*:* Despite the vertiginous levels of growth and promise, serious charges to industry profitability have recently emerged: (a) Consolidation: From the nearly 60 cellular companies in US, virtually all of them are now bankrupt, bought out, or struggling with heavy debts. Only six big players now account for 80% of the wireless pie. We have seen such consolidation in India too, where there are four major players (b) Growth: With 1.2 billion subscribers, India is currently the second-largest telecom market in the world and has seen rapid expansion over the past years. The industry has increased primarily due to favourable regulatory conditions, low prices, increased accessibility, and the introduction of Mobile Number Portability (MNP). The telecom sector is set to grow at a Compound Annual Growth Rate (CAGR) of 9.4% from 2020 to 2025. However, with a CAGR of 15.9% throughout the forecast period, the smartphone industry in India will have the fastest growth. However, as the growth increases, so have the competition; (c) Competition: As an obvious result (and to the consumer's delight), firms engaged in a devastating price war that not only eroded revenue growth but also endangered their ability to meet their titanic debts. (d) Customer Strategy: The industry paradigm has arguably changed from one of "make big networks, get customers" to "make new services, please customers." In short, the industry has moved from an acquisition orientation to a retention orientation.

The Elusive Customer*:* Until now, firms have been able to acquire customers without much effort. Demand for wireless services has been such that if a customer decided to drop his service and switch to

another carrier, another new customer was right behind him. The priority was to maintain the customer acquisition rate high, often at the expense of customer retention. But this situation has changed. As the well of wireless subscribers has begun to run dry, churn – the customer's decision to end the relationship and switch to another company – has become a major concern. Last year the industry average churn rate was 20% - 25% annually, which translates to approximately 2% churn per month. This means that companies lose 2% of their customers every month. Third quarter, 2001 (the data is a bit old, but the pattern still is similar), statistics show annual churn rates in an even higher range, 28% - 46% annual churn.

Churn rates for major carriers - Q3 2001



Source: *Telephony Online*, 2002

The reasons for the high level of churn are: (a) number of companies, (b) the similarity of their offerings, and (c) the cheap prices/perceived quality of service. In fact, the biggest current barrier to churn – the lack of phone number portability – has been removed and people can churn without changing their phone numbers. Companies are now beginning to realize just how important customer retention is. In fact, one study finds that "the top six US wireless carriers would have saved $207 million if they had retained an additional 5% of customers open to incentives but who switched plans in the past year" (Reuters 2002). Over the next years, the industry's biggest marketing challenge will be to control churn rates by identifying those customers who are most likely to leave and taking appropriate steps to retain them. The first step therefore is predicting churn likelihood at the customer level.

Data Description

The data provided have generously been provided to the Center by a major wireless carrier. The data are organized into three data files: Calibration, Current Score Data, and Future Score Data.

	Calibration	Current Score Data	Future Score Data
Sample Size	100,000	51,306	100,462
# of Predictor Variables	171	171	171
Churn Indicator	Yes	No	No

Customer ID	1,000,001 – 1,100,000	2,000,001 – 2,051,306	3,000,001 – 3,100,462

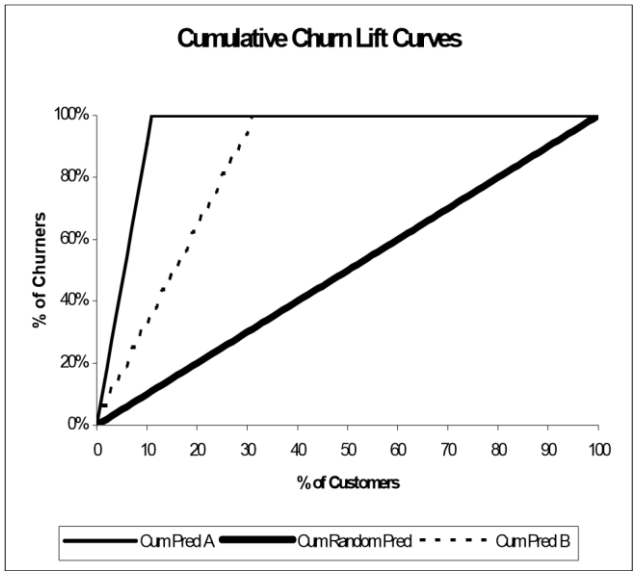
The Calibration Data contain the "dependent variable" – churn – as well as several potential predictors. The Current and Future Score Data contain the predictors but not churn. You are expected to develop your models on the calibration data and use these models to predict for the Current and Future Score Data, once you have evaluated their performance parameters.

The "Data Documentation" spreadsheet provides detailed descriptions of all the variables. The predictors include three types of variables: *behavioral data* such as minutes of use, revenue, handset equipment; *company interaction data* such as customer calls into the customer service center, and *customer household demographics*.

Customers were selected as follows: mature customers, customers who were with the company for at least six months, were sampled during July, September, November, and December of 2001. For each customer, predictor variables were calculated based on the previous four months. Churn was then calculated based on whether the customer left the company during the period 31-60 days after the customer was originally sampled. The one-month treatment lag between sampling and observed churn was for the practical concern that in any application, a few weeks would be needed to score the customer and implement any proactive actions.

The actual percentage of customers who churn in a given month is approximately 1.8%. However, churners were over sampled when creating the Calibration sample to create a roughly 50-50 split between churners and non-churners (the exact number is 49,562 churners and 50,438 non-churners). Over sampling was not undertaken in creating the Current Score and Future Score validation samples. This is to provide a more realistic predictive test. The Current Score data contain a different set of customers from the Calibration data, but selected at the same point in time. The Future Score data contain a different set of customers selected at a future point in time.

In addition to the regular measures, you will have to calculate two additional measures of predictive accuracy for each submitted data file – Top Decile Lift and Gini Coefficient. Top Decile Lift measures whether the 10% of customers predicted most likely to churn actually churn.



The Gini Coefficient measures predictive accuracy across the entire set of customers, not just the top 10%. The Gini Coefficient is used in economics to measure phenomena such as income inequality (Wolff 2002; Sydsaeter and Hammond 1995). In database marketing, the Gini Coefficient works off the "Cumulative Lift Curves," shown in the figure to the right. A Cumulative Lift Curve plots the top x% predicted customers versus the percentage of churners accounted for by these customers. For example, in the figure, the 10% of customers predicted most

likely to churn by Method B account for 31.3% of all churners. The top 20% predicted customers account for 62.5% of all churners. That is better than random prediction (shown by the Cum Random line), where the top 10% would account for 10% of churners, and the top 20% would account for 20% of churners.

The Gini Coefficient is the area between a method's cumulative lift curve and the random lift curve. Technically, it should be calculated as an integral (Sydsaeter and Hammond 1995) but we will approximate it by a numerical measure (Alker 1965; Statistics.Com, 2002) since we have a finite number of customers and no closed form formula for the cumulative lift curve for a given method.

The formula we use is:

$$Gini = \left(\frac{2}{n}\right)\sum_{i=1}^{n}(v_i - \hat{v}_i)$$

where:

n = number of customers,

v_i = % of churners who have predicted probability of churn equal to or higher than customer *i*,

$\hat{v}_i$ = % of customers who have predicted probability of churn equal to or higher than customer i

v_i is the height of the method's cumulative lift curve at the ith most likely predicted-to-churn customer, and $\hat{v}_i$ is the height of the random cumulative lift curve. The difference provides the "length" for calculating the area between the random and method prediction curves. The term 1/n approximates the "width" on the x-axis. The Gini Coefficient sums these lengths-times-widths across customers, providing an approximation to the area between the method's lift curve and the random lift curve. The calculation is multiplied by "2" to ensure that the maximum possible Gini Coefficient is 1[2].

The Gini Coefficient for Method A in the above figure is 0.84; the Gini for Method B is 0.69. Random prediction will achieve a Gini of 0 (as seen in the formula above since for random prediction, $\hat{v}_i = v_i$, and higher Gini will correspond to more separation between the method's lift curve and random, which means better prediction.

You, as a group, are required to upload a Word/pdf document describing the process that you followed, the results that you got, the plots that you have developed, interpretation of those in terms of your understanding, and finally the usefulness of the model to the telecommunications company. In addition, you need to upload a R script file which will run all the codes that you would have used to develop the model. If you use any other tools like Tableau or Excel, you are required to upload those files too. In addition, you are required to upload a video presentation, detailing out the process that you had followed, and the outcome thereof, including the interpretation of various results that you would have got, and your final recommendation of the model. This video should not be more than 15 minutes duration.

The evaluation will be based on the detailed explanation and interpretation incorporated in the Word Document, and the video presentation. In addition, you will be given due credit for concepts beyond what was discussed in the classes as incorporated by you in the task.

Your submissions will be checked for plagiarism, and if found to be plagiarized, will be heavily penalized.

All the best